HHS Public Access

Author manuscript

J Immunol. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

J Immunol. 2019 September 01; 203(5): 1252–1264. doi:10.4049/jimmunol.1801615.

Clustered mutations at the murine and human IgH locus exhibit significant linkage consistent with templated mutagenesis

Gordon A. Dale[1], **Daniel J. Wilkins**[1], **Caitlin D. Bohannon**[1], **Dario Dilernia**[1], **Eric Hunter**[1], **Trevor Bedford**[2], **Rustom Antia**[3], **Ignacio Sanz**[4], **Joshy Jacob**[1,*]

[1]Emory Vaccine Center, Yerkes National Primate Center, Emory University, 954 Gatewood Road, Atlanta, GA 30329 USA.

[2]Department of Epidemiology, School of Public Health, University of Washington, 1959 NE Pacific Street, Seattle, WA 98195 USA.

[3]Department of Biology, Emory University, 1510 Clifton Road, Atlanta, GA 30322 USA.

[4]Lowance Center for Human Immunology, Department of Medicine, Emory University, 615 Michael Street, Atlanta, GA 30322 USA.

Abstract

Somatic hypermutation generates a myriad of antibody mutants in antigen-specific B cells, from which high-affinity mutants are selected. Chickens, sheep, and rabbits use non-templated point mutations and templated mutations via gene conversion to diversify their expressed Ig loci, while mice and humans rely solely on untemplated somatic point mutations. Here, we demonstrate that in addition to untemplated point mutations, templated mutagenesis readily occurs at the murine and human Ig loci. We provide two distinct lines of evidence that are not explained by the Neuberger model of somatic hypermutation: (1) Across multiple data sets there is significant linkage disequilibrium between individual mutations, especially among close mutations. (2) Among those mutations, those less than 8bp apart are significantly more likely to match micro-homologous regions in the IgHV repertoire than predicted by the mutation profiles of somatic hypermutation. Together, this supports the role of templated mutagenesis during somatic diversification of antigen-activated B cells.

Keywords

Gene conversion; somatic hypermutation; humoral immunity; germinal center; B cell

Introduction

The ability of the humoral immune system to respond with high affinity to a vast assortment of antigens is critically dependent on the linked processes of somatic hypermutation and affinity-driven selection. The former generates somatically-mutated antibody substrates, from which clones with higher affinity than their peers can be selected to undergo further

*Correspondence to jjacob3@emory.edu.

somatic mutation. This process repeats itself until there is an emergence of a high affinity clone, specific for an antigen.

Evolutionarily, this strategy of somatic diversification and selection, emerged approximately 500 million years ago in vertebrates (1). In these earliest humoral responses, diversification occurred using a combination of a cytidine deaminase in tandem with gene conversion (2). Gene conversion is a mechanism of DNA repair wherein a highly homologous sequence is used as a template to repair a damaged region resulting in the copying of the template sequence into the damaged region. Further along the vertebrate evolutionary tree, the strategy of utilizing a cytidine deaminase to generate diversity is preserved throughout Mammalia (mammals) and Aves (birds). Chickens are known to rely heavily on activation induced cytidine deaminase (AID) and gene conversion. Conversely, mammals are canonically split, with animals such as rabbits (3), cattle (4), and sheep (5) being known to utilize gene conversion, whereas mice and humans are thought to primarily rely on AID and other nontemplated mutations generated from downstream processing of AID activity (6).

Here, we present evidence that somatic hypermutation in murine and human B cells utilizes a gene conversion-like mechanism, referred to hereafter as templated mutagenesis, to generate somatic variants. We observe templated mutagenesis in murine germinal center B cells as well as in terminally differentiated plasma cells. This observation is shared in human $IgM^+/IgG^+/IgA^+$ $CD138^+$ plasmablasts. We also observe linkage disequilibrium between mutations that is inversely related to genetic distance between those mutations at lengths <100bp.

Templated mutagenesis utilizes donors from variable segments 5’ to the rearranged VDJ in both mice and humans as well as from those on the other allele. Lastly, we find that non-immunoglobulin sequences placed at the IgH locus, by transgenes in mice and insertions in humans, mutate such that they share microhomology with tracts from the IgHV repertoire. Taken together, our studies demonstrate a role for templated mutagenesis during somatic hypermutation of murine and human B cells.

Methods

Contact for Reagent and Resource Sharing

Further information and requests for reagents should be directed to the corresponding author Joshy Jacob (jjacob3@emory.edu)

Experimental Model and Subject Details

Mice

C57BL/6 mice and CB6F1/J (C57BL/6 x BALB/cJ F1 hybrid) mice were purchased from The Jackson Laboratory. All mice were maintained in a specific pathogen-free facility in accordance with the institutional guidelines of The Animal Care and Use Committee at Emory University.

Method Details

Immunizations

Cohorts of female C57BL/6 or CB6F1/J mice were immunized intraperitoneally with 50μg hydroxyl-3-nitrophenylacetyl-chicken-γ-globulin (NP_{22}CGG) (Biosearch Technologies) with 50μL alum in PBS for a total volume of 200μL. Spleens and/or bone marrow were collected at time points described post-immunization.

High-throughput heavy chain sequencing

Total plasma cells ($CD138^+B220^-$), germinal center B cells ($CD19^+GL7^+$), or bone marrow B cells ($B220^+CD138^-CD19^+CD25^+IgM^-CD43^-$) were isolated via cell sorting from the spleens and bone marrow, respectively, of mice at 30 days post-immunization with NPCGG. Human plasmablasts ($CD19^+IgD^-CD27^+CD38^+CD138^+$) were sorted from peripheral blood of a healthy donor. Lysates were amplified and sequenced as previously reported (7, 8). Raw sequence data was then analyzed for mutations using The International Immunogenetics Information System® HighV-QUEST (IMGT.org) (9). Descriptive statistics for the sequence data can be found in Suppl. Table I, Part A. Sequences used in the analyses presented here can be found in Suppl. Table I, Part B.

Linkage Disequilibrium Plots

Plots of linkage disequilibrium between mutations were generated either with a custom python script, titled LD-analysis or a Matlab implementation of the correlation based tests described in Zaykin et al. (10).

The python script, LD-analysis, identifies major and minor alleles of polymorphic (mutated) sites from sequences grouped by IgHV gene usage. Only biallelic sites were considered. Major alleles are defined as those that are most common for a given site. Minor alleles are defined as the second most frequent alleles that also occur at frequency of ≥ 10% of observed sequences. From the pool of alleles for each position, haplotypes are generated. A chi-squared value is calculated for each haplotype and is subsequently used to generate a squared correlation coefficient (r^2) value that is reported for each haplotype in each LD plot.

The correlation-based tests were utilized for data sets where the restriction of data analysis to only biallelic sites was prohibitive with respect to number of haplotypes observed. In these cases, we report the maximum r^2 for a given haplotype pair.

Data presented is summed from all observed IgHV genes in each experiment, unless otherwise stated.

PolyMotifFinder and RandomCheck

To identify the number of potential templated mutagenesis events that could have contributed to observed somatic mutations, we generated a script with three objectives: 1) to identify mutations in sequences, 2) create motifs that include the mutated site, and 3) query this against the reference sequences. PolyMotifFinder is a script developed in Matlab (v.R2017b) for the purpose of identifying k-mer matches between raw sequence and

reference sequence datasets and studying their distribution and frequency (Figure S1). As inputs, PolyMotifFinder uses FASTA formatted files of the raw sequence data, an alignment of the raw sequence data to the germline sequence, and a series of unaligned reference sequences, in this case generated from the IMGT database (9). The script first identifies the positions where the aligned raw sequence and germline sequence exhibit mismatches, which is stored as a matrix of mutation positions. Then, the script identifies k-mer substrings from the unaligned raw sequence that incorporate two or more polymorphic positions that do not match the unaligned IMGT germline sequence (i.e. are mutations). Pulling the k-mers from unaligned sequences prevents gaps from indicating false mismatch polymorphisms when comparing to the IMGT sequences. The k-mer substrings are then quired against the IMGT reference sequences and if there is a matching IMGT reference sequence, the k-mer's coordinates are annotated in a matrix to indicate the length of the match. This matrix is compared to the matrix of mutation positions for each respective sequence. We tally the number of mutation positions that have a corresponding k-mer match and divide that by the number of mutations for a gene conversion (GC) coverage value for each sequence. This value is utilized below in tandem with the RandomCheck script.

The Matlab script RandomCheck was developed to produce a baseline for motif matching to compare the GC coverage results of PolyMotifFinder against. As input, the script takes the same data set as a PolyMotifFinder run. It identifies two or more mutations within one k-mer and randomizes the mutations, either keeping the polymorphism the same or changing it to another non-germline base. The randomization process is done such that it conforms to the base pair substitution profiles extensively reported to be characteristic of somatic hypermutation (Figures 3G and 4B) and is based on data reported from Longo et al. (human) and Maul et al. (mice) (11, 12). For example, to analyze a murine data set, a T→A mutation at position 50 of a given sequence is given a 27.8% chance of remaining a T→A mutation, a 55.6% chance to change to a T→C mutation, and a 16.7% chance to change to a T→G mutation. This new sequence is run through the same process as PolyMotifFinder, identifying how often the motifs match with somatic hypermutation-modeled combinations at the same positions as a real dataset. The GC coverage for each of these sequences is calculated and stored. This process is repeated with new somatic hypermutation modeled mutations 100 or 1000 times with the same k for direct comparison of results to the associated PolyMotifFinder run. For each sequence, the PolyMotifFinder GC-coverage is compared to the respective sequence's GC-coverage based on somatic hypermutation modeling, generating a Z-score.

Quantification and Statistical Analysis

Linkage disequilibrium analyses used for IgM plasma cells were conducted using correlation based tests described in Zaykin, Pudovkin, and Weir (10). Nucleotide positions that correspond to a putative gene conversion event were identified and used to calculate a R^2 value between all nucleotide positions that belong to a candidate gene conversion tract. For example, in Figure 1A, we analyzed nucleotide positions 25, 33, and 36 – which correspond to the A-----A--G tract from 1–53. Each pair of positions was tested for linkage. In this case, the statistic would compute the correlation between position 25 and 33, 25 and 36, and 33 and 36. These R^2 values would then be assessed for significance by permutating the alleles

of position 25, 33, and 36 and subsequently generating an R^2 value for each permutation. The original R^2 values are compared to the population of permutational R^2 values to generate a p-value. Exact (permutational) p-values are reported and are based on the value and distribution of R^2. 19,999 permutations were conducted for each analysis. Software was obtained from http://www.niehs.nih.gov/research/resources/software/biostatistics/rxc/index.cfm.

Stouffer's Z-method was used to determine statistical significance of gene conversion (GC) coverage as determined by PolyMotifFinder. Stouffer's Z is defined by:

$$Z_s = \frac{\sum_{i=1}^{n} Z_i}{\sqrt{n}}$$

where Z_S is Stouffer's Z value, n is the number of Z scores in the analysis, i is the i-th Z score out of a total of n Z scores. Unaltered somatically-mutated sequences were used to identify GC coverage and were compared to each sequence's respective population of GC coverage of altered somatically-mutated sequences generated by RandomCheck. Altered sequences had the identity of mutations changed to any nucleotide that was not germline but retained the mutations in their respective positions for each sequence as described above. GC coverage populations of altered sequences were determined for 100–1,000 iterations of mutation changes. The mean and standard deviation of each altered GC-coverage population was used to compute a Z-score for the respective unaltered GC-coverage for that sequence. Stouffer's Z method of meta-analysis was used to determine if the GC-coverage of the sequence sets is significantly different from that of modeled somatic hypermutation.

Stouffer's Z was combined between IgHVs of murine or human datasets via a weighted Stouffer's Z trend, defined as

$$Z_T = \frac{\sum_{S=1}^{n} w_S Z_S}{\sqrt{\sum_{S=1}^{n} w_S^2}}$$

where Z_T is Stouffer's Z trend, n is the number of Stouffer's Z scores used in the analysis, w is the weight of each Stouffer's Z score (defined as the number of Z scores for the corresponding Stouffer's Z divided by the total number of Z scores across all Stouffer's Z scores used in the analysis), S is the S-th Stouffer's Z score out of the maximum of n (13). Calculation of Stouffer's Z trend allows us to test whether there is an effect across different IgHV datasets and is weighted according to the number of Z-scores generated for each IgHV.

Data Availability

Sequences used in this study are deposited under SRA accession number PRJNA551000 and GEO accession number GSM2126020 and can be found at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA551000/ and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM2126020, respectively.

Results

Templated mutations occur in murine IgM$^+$ plasma cells

We have recently demonstrated the existence of antigen-specific, long-lived, somatically-mutated IgM plasma cells that have a high ratio of framework (FRW) mutations to that of the complementarity-determining regions (CDR) mutations (7). It is widely held that framework mutations are selected against since they are likely to disrupt the architecture of the antibody. Based on the current model of mutagenesis by point mutations, it is expected that FRW mutations occur more frequently than those in the CDRs and that these deleterious mutations are quickly selected against (14, 15). We posit that the FWR mutations in IgM plasma cells were non-deleterious and introduced via templated mutagenesis from other IgH FWR presumably through a gene conversion-like mechanism. To determine whether somatic mutation we observed in IgM$^+$ plasma cells was consistent with gene conversion, we intraperitoneally immunized cohorts of mice with 50μg of 4-hydroxy-3-nitrophenylacetyl labelled chicken-γ-globulin (NPCGG). We have previously shown these mutations to be dependent on the enzyme activation-induced cytidine deaminase (AID) (7), which is critical to both somatic hypermutation and gene conversion (16, 17). These mutations also result in the expected ~2:1 ratio of transitions and transversions, which is characteristic of somatic hypermutation (7). As previously reported, we observe a high frequency of replacement mutations in the framework region of IgM plasma cells as compared to those in the CDRs (Figure 1A). Further analysis of the ratio of replacement mutations to silent mutations in CDRs and FWRs revealed an enrichment of replacement mutations in framework regions 2 and 3 (Figure 1B).

We next aligned somatically-mutated IgHV 1–72 (V186.2) sequences, which canonically encode antibodies specific to the NP hapten (18–20), against the germline IgHV 1–72 - as well as the closely related sequences IgHV 1–55 and 1–53. Sequences shown in Figure 1C are representative of four individual mice. We find that many mutations within and between individual mice can be explained by gene conversion tracts (A…AAC…A) and (A…A…G) from the heavy chain variable genes segments IgHV 1–55 and 1–53, respectively (Figure 1C). IgHV 1–55 and −53 tracts occur in 5% and 20% of IgHV 1–72 sequences, respectively (Figure 1D). Together both tracts accounted for 25% of the mutation load observed in IgHV 1–72 (Figure 1E). Surprisingly, these tracts were observed in different clones within individuals and in different animals, strongly suggesting either concerted selective pressure in the germinal center or templated mutagenesis.

To examine whether these tracts were the result of selective pressure within the germinal center microenvironment, we analyzed linkage disequilibrium between silent and replacement mutations. Linkage disequilibrium (LD) is a measure that is used in genetics to test whether mutations are inherited together. In tracts from IgHV 1–53 (A…A…G), the first is a replacement mutation while the second and third are silent. As silent mutations are not selected for within the germinal center, it would follow that associations between silent mutations or between silent and replacement mutations would not be expected. A LD analysis found significant associations between silent mutations in tracts donated from IgHV1–53 ($p<0.001$). Additionally, the LD was significant between the replacement

mutation and both silent sites in the A…A…G tract from IgHV 1–53 ($p<0.001$). All mutations analyzed in tracts donated by IgHV1–55 (A…AAC…A) were replacement mutations and therefore could only be analyzed for linkage disequilibrium between replacements. However, analysis did reveal significant associations between the set of replacement mutations ($p<0.001$). Taken together, this suggests that IgM+ plasma cells do undergo somatic mutation consistent with templated mutagenesis.

Though unlikely, it is formally possible that the mutations could be due to PCR on bulk B cell population (21). To rule this out we analyzed the IgHV 1–72 sequences as reported by Tas et al. who sequenced sorted, single cells (22). We found that these single-cell derived IgHV 1–72 sequences possessed the replacement "A" and silent "A" mutations of the IgHV 1–53 tract in two distinct clones (Figure 1F). We also observed a multitude of clusters of mutations that share regions of varying microhomology with other IgHV genes, as opposed to the macro-homology observed between the IgHV 1–72 sequence and the germline IgHV 1–55 and −53 donors.

To further validate this, we sorted single plasma cells into 96 well plates and carried out PCR amplification for the gene rearrangements from Ig heavy and light chains. We obtained a total of 489 IgHV sequences but did not find any IgHV 1–72 sequences (n=6) with the IgHV 1–55 or IgHV 1–53 mutation tract in this limited cohort of IgHV 1–72 sequences. However, we did obtain somatically-mutated IgHV 1–53 sequences as well as somatically-mutated IgKV sequences (Figure S1). As observed in our analysis of the sequences reported in Tas et al., we found that clusters of mutations shared microhomology with germline IgHV and IgKV genes. As we find regions of microhomology accounting for mutations in both data sets, this suggests that the mutations observed in IgM plasma cells is not due to bulk PCR error. In addition, this suggests that clusters of mutations may be due to templated mutagenesis.

Templated mutations occur in murine germinal center B cells

IgM plasma cells appeared to have mutations consistent with templated mutagenesis, but we reasoned that tracts would be difficult to identify in IgG plasma cells since they accrue a high mutation burden. Hence, we analyzed developing (day 12) germinal center B cells as they have just begun to initiate somatic mutation. As hypothesized, we observed long (9–68bp) tracts from IgHV genes 1–55 and 1–53 in each of the three mice in both IgM^+ or IgG^+ germinal center B cells (Figure 2 A-B). We also observed the expected profile of somatic hypermutation, with transition mutations occurring twice as frequently than transversion mutations (Figure 2C-D). Long templated tracts of mutations occurred more frequently and accounted for more mutations in IgM germinal center B cells than in those that have switched to IgG (Figure 2E-G). To ensure that this pattern of mutation was not limited to the NPCGG response, we also analyzed Peyer's patch germinal center B cells which respond to gut flora antigens (Figure 1G) and identified tracts demonstrating that templated mutagenesis is not restricted to the hapten-carrier model.

It is possible that the clusters of mutations could have been introduced as independent point mutations as opposed to tracts. To distinguish between the two, we analyzed linkage disequilibrium of IgHV 1–72 sequences from IgM^+ and IgG^+ germinal center B cells from

day 12 along with all mutated sequences (n=11,456) from day 10, 12, and 14 NPCGG-induced splenic germinal centers. The data for all IgHV genes was then plotted as a function of distance between mutations (Figure 2H-J). Strikingly, we find high levels of linkage at lengths <50bp with multiple instances of complete linkage at these lengths as measured by the correlation coefficient r^2, where r^2=LD. Furthermore, a LOESS regression demonstrates that linkage decreases with increasing distance between mutations. To further validate this result, we reanalyzed data from Kuraoka et al. (23) who sequenced individual antigen-specific germinal center B cells from mice immunized 8 or 16 days earlier with either recombinant *Bacillus anthracis* protective antigen (rPA) or influenza hemagglutinin (rHA). Ig sequences from rPA at day 8 and day 16 as well as rHA at day 16 demonstrated the same LD phenomenon, suggesting that increasing LD at lengths <100bp was independent of the type of antigen used (Figure 2K-M). We also analyzed the CGG-specific sequences obtained from Tas et al. (22), which also demonstrates the same pattern observed in the Kuraoka sequences as well as those presented here (Figure 2N). Lastly, we analyzed somatically-mutated rabbit heavy and chicken light chain sequences (24–28), which undergo gene conversion predominantly and untemplated point mutation to a lesser extent and we found increasing LD with decreasing distance between mutations (Figure 2O-P).

It is known that AID preferentially targets specific motifs for hypermutation, thus we analyzed whether the observed increase in linkage was due to linked activity at nearby AID hotspots. We compared mutations at the canonical AID hotspot WRC/GYW to determine if there is an associated increase in linkage at mutations in close proximity. We analyzed the data sets from Tas et al., Kuraoka et al., as well as mutated rabbit and chicken sequences (22–28). We found that there were few WRC/GYW pairs that were mutated across data sets and that there was no consistent pattern to linkage between mutations at WRC/GYW sites (data not presented). An in-depth analysis of somatically mutated IgHV 1–72 sequences from splenic germinal centers from CB6F1/J mice following immunization with NPCGG (d12 p.i.) considered the recently reported hotspots (CRCY/RGYG and ATCT/AGAT) in addition to the canonical (WRC/GYW) revealed an increase in linkage equilibrium among mutations <100bp consistent with data presented in Figure 2E-M. However, the linkage between these sites represented a small fraction of the overall observed LD at distances <100bp, suggesting a minor contribution of linked AID deamination to observed LD between mutations at these sites (data not presented).

Although AID activity may be responsible for some of the observed LD that we observe across our data and others, we find the bulk of the LD occurs outside of regions that are AID hotspots suggesting that the paradigm of multiple AID-induced point mutations is not responsible for our LD observations. Though this data is not consistent with an AID-mediated pattern, it is consistent with gene conversion-like templated mutagenesis, as multiple mutations may be carried over together by the same templated event as evidenced from our analysis of somatically-mutated rabbit and chicken sequences.

Templated donor tracts primarily originate from 5’ upstream V gene segments but can also originate from the trans allele

In Fig.1 we showed that IgHV 1–72 rearrangements exhibit tracts from IgHV 1–55 and IgHV 1–53 gene segments. Interestingly, VDJ rearrangement of the IgHV 1–72 gene segment results in cis-deletion of the IgHV 1–55 and IgHV 1–53. Hence, we hypothesized that the only available template for these genes would be located on the trans allele. To test the contribution of the trans allele, we used a first filial generation BALB/cJ x C57BL/6J cross (CB6F1/J) for our studies. These mice express two different IgHV loci that express strain-specific alleles enabling identification of donor tracts from either allele into the expressed V_H gene.

To quickly identify tracts between IgHV genes from either the BALB/c or BL/6 locus we developed a computational script, PolyMotifFinder, that allows us to find small tracts of two or more mutations within 8bp donated from a set of reference sequences (Figure S1). We chose this strategy for PolyMotifFinder as we observed many mutations in close proximity that exhibited significant linkage disequilibrium, suggesting that these mutations may be due to a templated tract. We tested the robustness of the script to detect gene conversion tracts on the rabbit immunoglobulin sequences of Sehgal et al. (29) and found that this script detects >96% of mutations reported as gene conversion mutations (data not presented). PolyMotifFinder is supplemented with another script, RandomCheck (Figure S2), that is made to simulate the base pair substitution pattern of somatic hypermutation. Within each analysis, PolyMotifFinder generates a “Gene Conversion (GC) coverage” value for each sequence. That value is stored and RandomCheck assigns new mutations at the same position based on the profile of somatic hypermutation, followed by another GC coverage value calculation. The RandomCheck process is iterated either 100 or 1000 times to build a population, which the initial GC coverage is compared against, generating a Z score. The population of Z scores than then be combined into a single analysis using Stouffer’s Z method.

We immunized five CB6F1/J mice with NPCGG as above, sorted, and sequenced splenic germinal center B cells. The NP response differs in BALB/c and BL/6 mice, with different heavy chains encoding the antigen specific response, termed the NP^a and NP^b response and utilizes heavy chains IgHV 14–3 and IgHV 1–72, respectively (30). To determine whether tracts could originate from either cis or trans alleles, we analyzed IgHV 1–72 (NP^b) and IgHV 14–3 (NP^a) sequences (Figure 3A). In each of these sequence sets we observed increasing LD at decreasing genomic distance between mutations, as before (Figure 3B-C). Sequences were then analyzed with PolyMotifFinder and RandomCheck using the base pair substitution profiles generated from Maul et al. (12) (Figure 3D). Each set of sequences was compared to the germline IgHV gene segments in its entirety, the BALB/c specific IgHVs, the BL/6 IgHVs, 8-mer motifs specific to only the BALB/c or BL/6 IgHV repertoire, as well as the set of 8-mers not represented in any IgHV (negative control). All germline IgHV sequences were obtained via IMGT(9). In both IgHV 1–72 and IgHV 14–3, all groups were significant against the negative control. We also find an average Stouffer’s Z value of approximately five in both sets when the sequences are compared to the entire IgHV repertoire, which suggests that templated mutagenesis is occurring over the background of

modeled somatic hypermutation (Figure 3E-F). Both the BL/6 and BALB/c specific motifs had Stouffer's Z scores around zero, suggesting that there was not significant matching to either of these sequence sets over what is predicted by somatic hypermutation. Conversely both the BL/6 and BALB/c IgHV sets produced average Stouffer's Z scores comparable to that of the entire IgHV repertoire, suggesting that templated mutagenesis is utilizing motifs shared between the BL/6 and BALB/c IgHV gene segments but did not resolve whether mutations were occurring in *cis* or in *trans*.

Since F1 mice possess a single copy of the BL/6 IgH locus, templated mutation that occurs in cis is predicted to be directional, as IgHV genes lost during VDJ recombination are not preserved on the opposite locus. Therefore, for any given IgHV rearrangement, we could analyze the donors located 5' as occurring in cis, and any matching to downstream donors is inferred to occur in trans with homologous motifs located on the BALB/c IgH locus. Although this analysis is possible with the BL/6 IgH locus, as the locus map is known, it was not possible to do this analysis on the BALB/c locus as there remains some ambiguity on its organization(31). As such, we limited our analysis to the BL/6 locus alone. We performed deep sequencing of the antibody repertoire present in the germinal centers of F1 mice immunized with NPCGG. For each mouse, each BL/6 IgHV gene is shown in the order in which the locus is organized. Across mice, 5' donors are preferentially utilized (Figure 3G) ($p<0.0001$). However, analysis of the 3' donors reveals that there is significant enrichment of 3' donor matching as compared to the somatic hypermutation null, suggesting that templated mutation can also occur in trans, albeit with a much lower frequency (Figure 3H). Interestingly, we observed that IgHV genes located towards the 5' end of the IgH locus displayed a consistent enrichment of 3' matching, suggesting the lack of available templates forces templated mutation to occur in trans, as observed in IgHV 1–72 sequences earlier.

Templated mutation occurs in human plasmablasts

We next sought to investigate if somatic hypermutation in human B cells also occurred due to templated mutagenesis, as we have observed in mice. To do so, we analyzed circulating plasmablasts isolated from the blood of a healthy human donor. Sequences from these cells were generated using ultra-high throughput single cell sequencing as described in DeKosky et al. (32). As shown in other sequence sets, linkage disequilibrium increases at mutation distances less than 100 bp (Figure 4A). To determine whether this observation was due templated mutagenesis, we analyzed each antibody rearrangement in our sample and compared whether there was significant matching to germline human IgHV gene segments as defined by IMGT (9), or to other 8-mers not present in the germline IgHV gene segments using the human SHM base pair substitution profiles derived from Longo et al (11) (Figure 4B). There is significant matching to the IgHV germline across all IgHVs sequenced (Stouffer's Z Trend: 4.33, $p<0.0001$) as compared to matching motifs not present in the IgHV germline gene segments (Stouffer's Z Trend: −35.62, $p=1$) (Figure 4C). When analyzed across isotypes (IgM/IgG/IgA) we found a similar pattern in each group (Stouffer's Z trend: 1.5235, 1.7664, 1.5922, respectively) compared to the negative control of 8-mers not present in the human IgHV repertoire (Stouffer's Z trend: −17.6049, −12.3690, −16.1749, respectively) (Figure 4D-F). Lastly, we hypothesized that templated mutagenesis would occur primarily using 5' donors as demonstrated in mice. Analysis via

PolyMotifFinder and RandomCheck demonstrated statistically significant matching to upstream 5' donors in human plasmablasts as compared to those downstream (Stouffer's Z trend: 4.04 and 1.84, respectively, $p<0.05$) (Figure 4G). Together, these results suggest a similar pattern of somatic hypermutation, consistent with templated mutagenesis, in diversification of antigen-activated human B cells.

Templated mutations occur in non-Ig genes inserted into the IgH locus

It has been shown that transgenes inserted into the murine Ig loci as well as insertion of LAIR gene into the human Ig loci undergo somatic mutation (33, 34). Thus, we next sought to examine the patterns of mutation of non-immunoglobulin genes that undergo somatic hypermutation at the IgH locus in mice and humans. To do this, we compared the somatically-mutated LAIR1 insert of the V-LAIR1-DJ antibodies reported in Tan et al. (33) (Figure 5A) against the human IgHV repertoire using PolyMotifFinder and RandomCheck scripts. We also utilized the transgenes β-globin and GPT placed at the passenger IgH allele of mice, where they are not subject to selective pressure, against the murine IgHV repertoire (35) (Figure 5B). As before, we also compared these sequences to 8-mers not present in the respective IgHV repertoires (negative control). In humans, somatically-mutated LAIR1 preferentially matches the IgHV repertoire as compared to the negative control ($p=0.004$) (Figure 5D). With regards to mice, Alt and colleagues published an elegant paper (35) wherein they generated genetically engineered mice which carried passenger transgenes, β-globin and GPT introduced into the Ig loci. These transgenes cannot undergo selection as they were engineered to be transcribed but not translated. Interestingly, the passenger transgenes β-globin and GPT displayed the same pattern as LAIR1 and are statistically significant in matching motifs present in the murine IgHV repertoire ($p=0.018$ and $p=0.015$, respectively) (Figure 5D-F). For LAIR1, β-globin, and GPT, the negative control was not significant ($p= 0.972$, $p= 0.995$, $p= 0.998$, respectively). This suggests that patterns of mutations in these three datasets (one human and two murine) are consistent with the templated mutagenesis from the IgHV gene segments as the pattern of somatic hypermutation statistically results in motifs matching those in the IgHV repertoire.

To demonstrate that this effect is directly due to templated mutagenesis acting on both IgHV and these non-immunoglobin sequences, we conducted a second analysis in which we first analyze somatically-mutated IgHV sequences and extract those motifs with two or more mutations that match an 8-mer in the IgHV repertoire. These matched motifs are then used as the reference against which somatically-mutated non-immunoglobulin sequences are compared (Figure 5C). We reasoned that the templates (motfis) used for templated mutagenesis of the IgHV gene segments would be the same as those used for the non-immunoglobulin sequences if templated mutagenesis were occurring. Thus, we hypothesized that we would observe a stronger effect when analyzing the somatically-mutated non-immunoglobulin sequences against these enriched motifs, since we only selected the motifs used during templated mutagenesis and excluded those that were not. As hypothesized, Stouffer's Z trend generated from the enriched pool rises for each sequence set when compared against the IgHV references ($p=0.023$) (Figure 5D-F). Further analysis has also shown that the enrichment effect on Stouffer's Z trend is not limited to enriching for motifs from somatically-mutated IgHV genes with comparison to mutated non-immunoglobulin

sequences but also occurs if the enrichment occurs from somatically-mutated non-immunoglobulin genes and mutated IgHVs are compared to that enriched pool (Figure S3). These results demonstrate that the pattern of templated mutation is consistent between somatically-mutated IgHV genes and exogenous genes inserted into the IgH locus and that such effect is not due to the intrinsic activity of canonical somatic hypermutation.

Lastly, we sought to quantify the contribution of templated mutagenesis to somatic hypermutation. In order to do so we conducted a conservative and limited-in-scope analysis that (1) considered mutations within 8bp of at least on other mutation, (2) only considered one instance of a given pair of mutations (to eliminate double counting within a dataset), and (3) did not factor clonality so as to remove any bias from clonal dynamics. This approach allows us to determine the likelihood that a given set of mutations in close proximity has a template in the IgHV segment repertoire. We analyzed human and murine sequence sets from Tas et al. (22), human plasmablasts, the LAIR1 gene segments (33, 34), as well as the passenger transgenes (β-globin and gpt) from Yeap et al. (35). We find that approximately 50–65% of unique mutations fulfill two conditions: 1) proximity to at least one other mutation within 8bp, and 2) there exists a template for those mutations in the IgHV germline gene segments (Figure 5I). Strikingly, this effect extends into the somatically mutated non-Ig sequences, LAIR1, gpt and β-globin, despite the lack of overt homology between these genes and the IgHV repertoire. Extrapolation of this data suggests that 3 out of every 5 mutations at the IgH locus are consistent with templated mutagenesis, whether the gene being mutated is an antibody gene segment or a non-immunoglobulin gene. This is especially significant for the passenger transgenes β-globin and gpt which are not subjected to selection and display the full spectrum of mutations as they occur. Together, this data suggests that templated mutation contributes heavily to mutation clusters, specifically those that are within 8bp of one another.

Discussion

In the present study we provide evidence of templated mutagenesis occurring during antigen-driven somatic diversification of human and murine B cells. We find that this is an intrinsic property of somatic hypermutation, at least as it occurs at the IgH locus in both species, as the non-immunoglobulin sequences examined are enriched for motifs that are exceedingly unlikely to have occurred by the effects of canonical somatic hypermutation. We suspect but cannot confirm within the scope of the current study, that the mechanism of templated mutagenesis is gene conversion. The repeated occurrence of IgHV repertoire motifs within somatically-mutated sequences is akin to that seen in other species such as the chicken, which relies heavily on acquisition of pseudogene motifs into the productive antibody rearrangement (36). However, unlike chickens, the pattern of templated mutagenesis in mice and humans is largely based on IgHV microhomology, as small fragments around pairs of somatic mutations disproportionately match germline IgHV motifs. This contrasts with typical reports of gene conversion in the literature which rely on large alignments of potential donors and recipient sequences to demonstrate that the mutations are templated. Further, reports of gene conversion in the literature rely on demonstrating an arbitrary number of mutations matching to a pseudogene sequence to establish a given gene conversion tract.

In this work, we address these two limitations by applying a microhomology and statistical approach to identifying potential gene conversion tracts that queries whether it was possible to use fragments of IgHV genes to reconstruct highly-mutated antibody sequences. This process was further extended in our script, which identifies if “neomotifs” – those generated after introduction of somatic mutations – are actually represented in the genetic repertoire of the IgHV loci and does so without regard to large alignments or total number of templated mutations in a tract. This has allowed us to identify potential gene conversion tracts that would have otherwise been excluded in a conventional analysis. Interestingly, gene conversion has been previously suggested to occur in both mice (37–40) and humans (41, 42), but has generally been regarded as an infrequent event or even thought to not occur at all (43).

Here, we argue in favor of the alternative, where gene conversion or a mechanism akin to it, is a frequent contributor of somatic mutations along the length of the IgHV gene. This is most evident by the ability of “neomotifs” present in non-selected, passenger transgenes *GPT* and β-globin (35) to (1) match murine IgHV reference sequences and (2) for those motifs that do match the IgHV repertoire, produce significant results for somatically-mutated IgHV genes (Figure S3). This demonstrates that the process of templated mutagenesis is a significant contributor to somatic hypermutation, else such an effect would not occur. Indeed, our conservative estimates of the contribution of templated mutagenesis to somatic hypermutation suggest that between 50–65% of unique mutations occur 1) in close proximity to another mutation and 2) have a putative donor sequence in the IgHV germline repertoire that explains both, or more, mutations in that cluster. That this frequency is observed even within the non-immunoglobulin passenger transgenes in Yeap et al (35) is striking because despite a lack of overt homology with the IgHV repertoire, and absence of selective pressure (Figure 5B), they remarkably have clusters of mutations that match templates in the IgHV repertoire. Coupled with the finding that templates used for diversifying the IgHV genes also serve as templates for diversifying these non-immunoglobulin genes, these findings strongly support the notion of a templated mechanism for the production of local micro-clusters (≤8bp) of mutations.

Direct additional evidence for this templated pattern of mutation, while rare, can be found in published literature. To illustrate, in the classic Nature paper, Weiss and colleagues (18) micro-dissected individual germinal center B cells, sequenced them and demonstrated that somatic mutants are generated intra-clonally within germinal centers. In Figure 2 of this paper, sequences GC24.I6 and GC24.I12 both contain a replacement G→A substitution at codon 9, and a silent G→A substitution at codon 10. This pair of substation mutations are seen in the IgHV 1–53 tracts presented in our manuscript. Examples of this tract in our manuscript can be seen across multiple clones in both IgM plasma cells, and in single cell sequenced germinal center B cells from the Victora Lab (22), depicted in Figures 1C and D respectively. The IgHV 1–55 tract appears less frequently in the literature but we have observed it in Jacob et al. (44), Figure 6. Sequence B17–2 possesses the three-nucleotide replacement GCA→ AAC in codon 34–35 that we observe in the IgM plasma cells sequences (Fig.1C; current manuscript).

As to why such tracts may seem rare in the literature – when Kelsoe and colleagues conducted these studies in 1991, there was a primary concern for PCR cross-over artifacts. At the time, single cell sequencing was best done by successful manual dissection and there was no obvious control for successful isolation of a single cell. Thus, when these studies were done, “PCR hybrids” that were found were eliminated from any further analysis. Interestingly, the occurrence of PCR hybrid generation was investigated by Kelsoe and colleagues (19), Figure 8. They found a steady increase in the formation of PCR hybrid products as the germinal center reaction progressed. At the time of the study, this effect was attributed to either DNA nicking or apoptotic DNA fragmentation – neither of which could be a possible explanation for our observations in the current manuscript since the IgM plasma cells are (1) successful emigrants from the germinal center reaction and are expected to have repaired any existing DNA lesions, (2) the IgM plasma cells were not undergoing apoptotic processes as sorted cells were stained with a live/dead stain (Annexin V) and were gated on live cells during sorting, and (3) our sequence data was generated from cDNA libraries and reflect the mRNA transcripts present in the IgM plasma cell population, thus excluding DNA damage as a potential explanation. These findings, however, are consistent with templated mutagenesis and the observations presented here.

One of the primary arguments against the occurrence of gene conversion during murine somatic mutagenesis was a series of studies by Bross et al. (45). These studies sought to elucidate the role of homologous recombination (gene conversion) during somatic hypermutation by analyzing mice deficient in RAD54. The primary function of RAD54 is to facilitate homologous recombination through branch migration (46). They found that the frequency and pattern of somatic hypermutation was unaltered despite the absence of RAD54 and concluded that there was no contribution of RAD54 to somatic hypermutation. In contrast, another study published around the same time by D’Avirro et al. (47) demonstrated a gene conversion-like phenomenon in IgH knock-in mice that were RAD54$^{-/-}$. They concluded that the gene conversion-like phenomenon was independent of RAD54, suggesting that templated mutagenesis could still readily occur in the absence of RAD54. This is further supported by studies in *S. cerevisiae* that demonstrate that disruption of the RAD54 gene produces a mild but not critical defect in mating-type switch, a process entirely dependent on gene conversion (48). Based on the results from these studies, it remains plausible that gene conversion-like events occur and can do so independently of RAD54.

The current paradigm of somatic hypermutation is the Neuberger model (49), which describes the many processing events that occur downstream of AID-mediated deamination of cytosine. While the Neuberger model accounts for many observations made regarding somatic hypermutation, the data presented here is not adequately explained under this paradigm. Foremost, we observe a replicable increase in linkage disequilibrium of somatic mutations as the distance between those mutations decreases. Secondly, we find that these pairs of somatic mutations in close proximity, with the highest linkage, disproportionately match the IgHV repertoire, regardless of whether the mutated sequence is an IgHV gene segment or a non-immunoglobulin gene. Together, both lines of evidence suggest that templated mutation is an actively-occurring process during somatic hypermutation. Such a mechanism carries significant implications for the generation of diversity during the humoral

immune response and further work should be aimed at elucidating the functional impact such a mechanism has on antibody affinity maturation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge members of the Jacob and Antia Laboratory for helpful discussions. We thank Leela Thomas for mouse colony management.

This work was supported by NIH grants U19 AI117891 and F30 AI124568

Findings:

Somatic mutations in close proximity exhibit high linkage disequilibrium

Pairs of somatic mutations within 8bp are templated from IgHV gene segments

airs of somatic mutations in non-Ig genes at the IgH locus are similarly templated

Works Cited

1. Cooper MD, and Herrin BR. 2010 How did our complex immune system evolve? Nat Rev Immunol 10: 2–3. [PubMed: 20039476]
2. Boehm T, McCurley N, Sutoh Y, Schorpp M, Kasahara M, and Cooper MD. 2012 VLR-based adaptive immunity. Annu Rev Immunol 30: 203–220. [PubMed: 22224775]
3. Becker RS, and Knight KL. 1990 Somatic diversification of immunoglobulin heavy chain VDJ genes: evidence for somatic gene conversion in rabbits. Cell 63: 987–997. [PubMed: 2124176]
4. Meyer A, Parng CL, Hansal SA, Osborne BA, and Goldsby RA. 1997 Immunoglobulin gene diversification in cattle. Int Rev Immunol 15: 165–183. [PubMed: 9222818]
5. Butler JE 1998 Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals. Rev Sci Tech 17: 43–70. [PubMed: 9638800]
6. Di Noia JM, and Neuberger MS. 2007 Molecular mechanisms of antibody somatic hypermutation. Annu Rev Biochem 76: 1–22. [PubMed: 17328676]
7. Bohannon C, Powers R, Satyabhama L, Cui A, Tipton C, Michaeli M, Skountzou I, Mittler RS, Kleinstein SH, Mehr R, Lee FE, Sanz I, and Jacob J. 2016 Long-lived antigen-induced IgM plasma cells demonstrate somatic mutations and contribute to long-term protection. Nat Commun 7: 11826. [PubMed: 27270306]
8. Tiller T, Busse CE, and Wardemann H. 2009 Cloning and expression of murine Ig genes from single B cells. J Immunol Methods 350: 183–193. [PubMed: 19716372]
9. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, Carillon E, Duvergey H, Houles A, Paysan-Lafosse T, Hadi-Saljoqi S, Sasorith S, Lefranc G, and Kossida S. 2015 IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. Nucleic Acids Res 43: D413–422. [PubMed: 25378316]
10. Zaykin DV, Pudovkin A, and Weir BS. 2008 Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics 180: 533–545. [PubMed: 18757931]
11. Longo NS, Lugar PL, Yavuz S, Zhang W, Krijger PH, Russ DE, Jima DD, Dave SS, Grammer AC, and Lipsky PE. 2009 Analysis of somatic hypermutation in X-linked hyper-IgM syndrome shows specific deficiencies in mutational targeting. Blood 113: 3706–3715. [PubMed: 19023113]
12. Maul RW, MacCarthy T, Frank EG, Donigan KA, McLenigan MP, Yang W, Saribasak H, Huston DE, Lange SS, Woodgate R, and Gearhart PJ. 2016 DNA polymerase iota functions in the generation of tandem mutations during somatic hypermutation of antibody genes. J Exp Med 213: 1675–1683. [PubMed: 27455952]

13. Zaykin DV 2011 Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J Evol Biol 24: 1836–1841. [PubMed: 21605215]
14. Liu YJ, Joshua DE, Williams GT, Smith CA, Gordon J, and MacLennan IC. 1989 Mechanism of antigen-driven selection in germinal centres. Nature 342: 929–931. [PubMed: 2594086]
15. Shlomchik MJ, Watts P, Weigert MG, and Litwin S. 1998 Clone: a Monte-Carlo computer simulation of B cell clonal expansion, somatic mutation, and antigen-driven selection. Curr Top Microbiol Immunol 229: 173–197. [PubMed: 9479855]
16. Arakawa H, Hauschild J, and Buerstedde JM. 2002 Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. Science 295: 1301–1306. [PubMed: 11847344]
17. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, and Honjo T. 2000 Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell 102: 553–563. [PubMed: 11007474]
18. Jacob J, Kelsoe G, Rajewsky K, and Weiss U. 1991 Intraclonal generation of antibody mutants in germinal centres. Nature 354: 389–392. [PubMed: 1956400]
19. Jacob J, Przylepa J, Miller C, and Kelsoe G. 1993 In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. III. The kinetics of V region mutation and selection in germinal center B cells. J Exp Med 178: 1293–1307. [PubMed: 8376935]
20. Toellner KM, Jenkinson WE, Taylor DR, Khan M, Sze DM, Sansom DM, Vinuesa CG, and MacLennan IC. 2002 Low-level hypermutation in T cell-independent germinal centers compared with high mutation rates associated with T cell-dependent germinal centers. J Exp Med 195: 383–389. [PubMed: 11828014]
21. De Semir D, and Aran JM. 2003 Misleading gene conversion frequencies due to a PCR artifact using small fragment homologous replacement. Oligonucleotides 13: 261–269. [PubMed: 15000840]
22. Tas JM, Mesin L, Pasqual G, Targ S, Jacobsen JT, Mano YM, Chen CS, Weill JC, Reynaud CA, Browne EP, Meyer-Hermann M, and Victora GD. 2016 Visualizing antibody affinity maturation in germinal centers. Science 351: 1048–1054. [PubMed: 26912368]
23. Kuraoka M, Schmidt AG, Nojima T, Feng F, Watanabe A, Kitamura D, Harrison SC, Kepler TB, and Kelsoe G. 2016 Complex Antigens Drive Permissive Clonal Selection in Germinal Centers. Immunity 44: 542–552. [PubMed: 26948373]
24. Schiaffella E, Sehgal D, Anderson AO, and Mage RG. 1999 Gene conversion and hypermutation during diversification of VH sequences in developing splenic germinal centers of immunized rabbits. J Immunol 162: 3984–3995. [PubMed: 10201919]
25. Sehgal D, Obiakor H, and Mage RG. 2002 Distinct clonal Ig diversification patterns in young appendix compared to antigen-specific splenic clones. J Immunol 168: 5424–5433. [PubMed: 12023335]
26. Winstead CR, Zhai SK, Sethupathi P, and Knight KL. 1999 Antigen-induced somatic diversification of rabbit IgH genes: gene conversion and point mutation. J Immunol 162: 6602–6612. [PubMed: 10352277]
27. Arakawa H, Kuma K, Yasuda M, Ekino S, Shimizu A, and Yamagishi H. 2002 Effect of environmental antigens on the Ig diversification and the selection of productive V-J joints in the bursa. J Immunol 169: 818–828. [PubMed: 12097385]
28. Mansikka A, Sandberg M, Lassila O, and Toivanen P. 1990 Rearrangement of immunoglobulin light chain genes in the chicken occurs prior to colonization of the embryonic bursa of Fabricius. Proc Natl Acad Sci U S A 87: 9416–9420. [PubMed: 2123557]
29. Sehgal D, Mage RG, and Schiaffella E. 1998 VH mutant rabbits lacking the VH1a2 gene develop a2+ B cells in the appendix by gene conversion-like alteration of a rearranged VH4 gene. J Immunol 160: 1246–1255. [PubMed: 9570541]
30. Loh DY, Bothwell AL, White-Scharf ME, Imanishi-Kari T, and Baltimore D. 1983 Molecular basis of a mouse strain-specific anti-hapten response. Cell 33: 85–93. [PubMed: 6432337]
31. Retter I, Chevillard C, Scharfe M, Conrad A, Hafner M, Im TH, Ludewig M, Nordsiek G, Severitt S, Thies S, Mauhar A, Blocker H, Muller W, and Riblet R. 2007 Sequence and characterization of

the Ig heavy chain constant and partial variable region of the mouse strain 129S1. J Immunol 179: 2419–2427. [PubMed: 17675503]

32. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, and Georgiou G. 2015 In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med 21: 86–91. [PubMed: 25501908]
33. Tan J, Pieper K, Piccoli L, Abdi A, Foglierini M, Geiger R, Tully CM, Jarrossay D, Ndungu FM, Wambua J, Bejon P, Fregni CS, Fernandez-Rodriguez B, Barbieri S, Bianchi S, Marsh K, Thathy V, Corti D, Sallusto F, Bull P, and Lanzavecchia A. 2016 A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. Nature 529: 105–109. [PubMed: 26700814]
34. Pieper K, Tan J, Piccoli L, Foglierini M, Barbieri S, Chen Y, Silacci-Fregni C, Wolf T, Jarrossay D, Anderle M, Abdi A, Ndungu FM, Doumbo OK, Traore B, Tran TM, Jongo S, Zenklusen I, Crompton PD, Daubenberger C, Bull PC, Sallusto F, and Lanzavecchia A. 2017 Public antibodies to malaria antigens generated by two LAIR1 insertion modalities. Nature 548: 597–601. [PubMed: 28847005]
35. Yeap LS, Hwang JK, Du Z, Meyers RM, Meng FL, Jakubauskaite A, Liu M, Mani V, Neuberg D, Kepler TB, Wang JH, and Alt FW. 2015 Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. Cell 163: 1124–1137. [PubMed: 26582132]
36. Reynaud CA, Anquez V, Grimal H, and Weill JC. 1987 A hyperconversion mechanism generates the chicken light chain preimmune repertoire. Cell 48: 379–388. [PubMed: 3100050]
37. Cumano A, and Rajewsky K. 1986 Clonal recruitment and somatic mutation in the generation of immunological memory to the hapten NP. EMBO J 5: 2459–2468. [PubMed: 2430792]
38. David V, Folk NL, and Maizels N. 1992 Germ line variable regions that match hypermutated sequences in genes encoding murine anti-hapten antibodies. Genetics 132: 799–811. [PubMed: 1468632]
39. D'Avirro N, Truong D, Xu B, and Selsing E. 2005 Sequence transfers between variable regions in a mouse antibody transgene can occur by gene conversion. J Immunol 175: 8133–8137. [PubMed: 16339551]
40. Xu B, and Selsing E. 1994 Analysis of sequence transfers resembling gene conversion in a mouse antibody transgene. Science 265: 1590–1593. [PubMed: 8079173]
41. Darlow JM, and Stott DI. 2006 Gene conversion in human rearranged immunoglobulin genes. Immunogenetics 58: 511–522. [PubMed: 16705406]
42. Sanz I 1991 Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. J Immunol 147: 1720–1729. [PubMed: 1908883]
43. Lavinder JJ, Hoi KH, Reddy ST, Wine Y, and Georgiou G. 2014 Systematic characterization and comparative analysis of the rabbit immunoglobulin repertoire. PLoS One 9: e101322. [PubMed: 24978027]
44. Jacob J, and Kelsoe G. 1992 In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. J Exp Med 176: 679–687. [PubMed: 1512536]
45. Bross L, Wesoly J, Buerstedde JM, Kanaar R, and Jacobs H. 2003 Somatic hypermutation does not require Rad54 and Rad54B-mediated homologous recombination. Eur J Immunol 33: 352–357. [PubMed: 12548566]
46. Mazin AV, Mazina OM, Bugreev DV, and Rossi MJ. 2010 Rad54, the motor of homologous recombination. DNA Repair (Amst) 9: 286–302. [PubMed: 20089461]
47. D'Avirro N, Truong D, Luong M, Kanaar R, and Selsing E. 2002 Gene conversion-like sequence transfers between transgenic antibody V genes are independent of RAD54. J Immunol 169: 3069–3075. [PubMed: 12218123]
48. Schmuckli-Maurer J, and Heyer WD. 1999 The Saccharomyces cerevisiae RAD54 gene is important but not essential for natural homothallic mating-type switching. Mol Gen Genet 260: 551–558. [PubMed: 9928934]
49. Maul RW, and Gearhart PJ. 2014 Refining the Neuberger model: Uracil processing by activated B cells. Eur J Immunol 44: 1913–1916. [PubMed: 24920531]

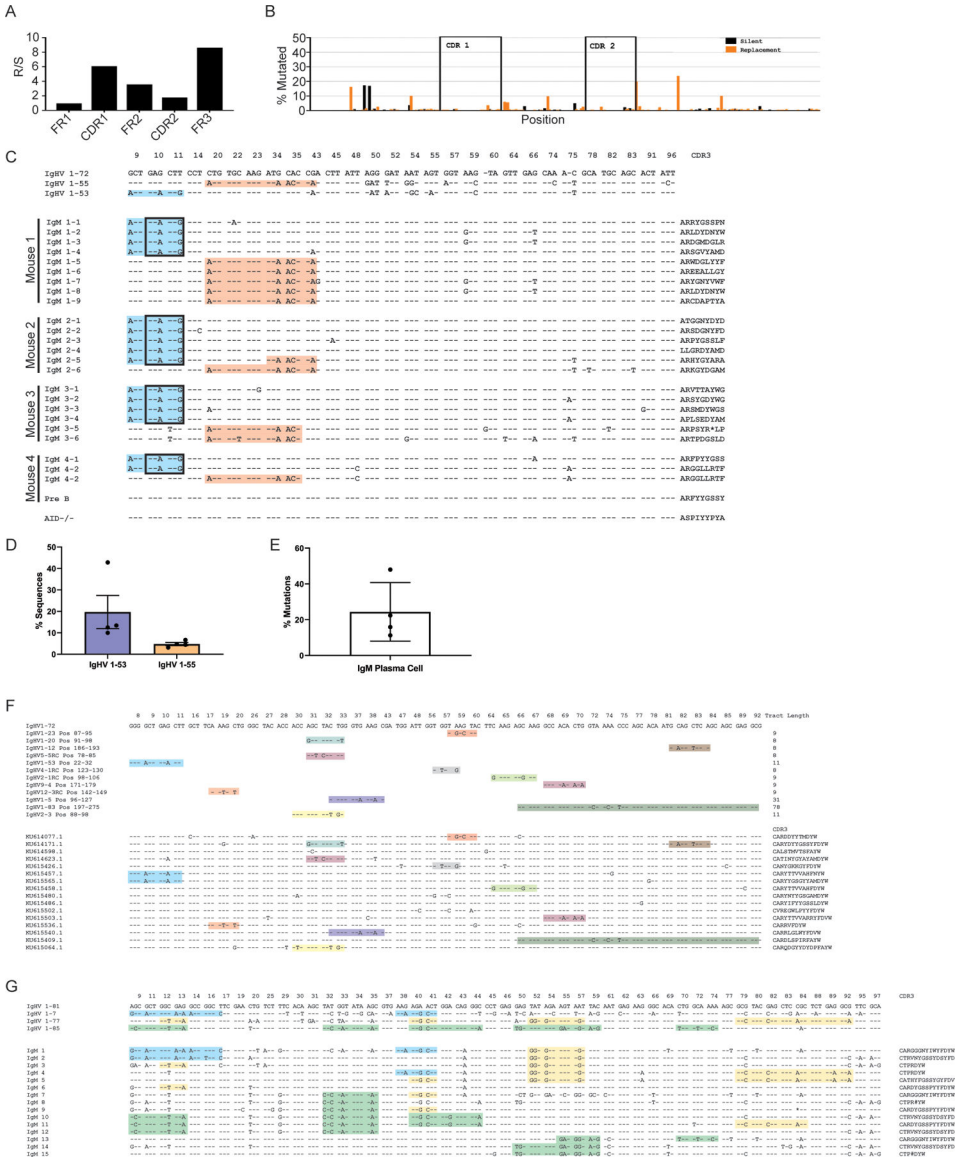

Fig. 1. Somatic hypermutation motifs in IgM plasma cells are shared between individual clones in individual mice.

(**A**) Overview of somatic hypermutation in IgM plasma cells. Shown are percent replacement and silent mutations per position along the IgHV 1–72 sequence. CDR 1 and 2 are highlighted with a box. Framework 1, 2, and 3 are intervening regions. (**B**) Replacement-to-silent mutation ratios for IgHV domains for IgM plasma cell sequences. (**C**) Nucleotide alignment of somatically-mutated IgM plasma cell sequences to germline IgHV 1–72, as well as putative donor germlines IgHV 1–55 and 1–53. Sequences are grouped according source mouse. Sequences shown in Fig. 1C are representative of four individual mice. CDR3 sequences are shown as amino acids. Only codons that differ in the alignment are shown. A representative pre-B cell sequence is shown as a sequencing control. Gene conversion tracts from IgHV 1–53 are colored blue, and IgHV 1–55 are colored orange. Boxed codons represent silent mutations. (**D**) Percent of IgM plasma cell sequences in Figure 1C that possess either the IgHV 1–53 gene conversion tract (blue) or the IgHV 1–55 gene

conversion tract (orange). (**E**) Shown are the percent of total mutations in the IgM plasma cell data set attributed to gene conversion with the IgHV 1–55 and IgHV 1–53 tracts. (**F**) Nucleotide alignment of somatically-mutated IgHV 1–72 sequences from Tas et al (2016). Sequences are presented as in **C**. Coordinates for donor IgHV fragments are shown on the left. RC denotes the noncoding strand. Tract length in bp is shown on the right for each tract. (**G**) Shown are somatically-mutated IgHV 1–81 sequences from germinal center B cells isolated from Peyer's patches of C57BL/6 mice. Data is presented as in **C**.

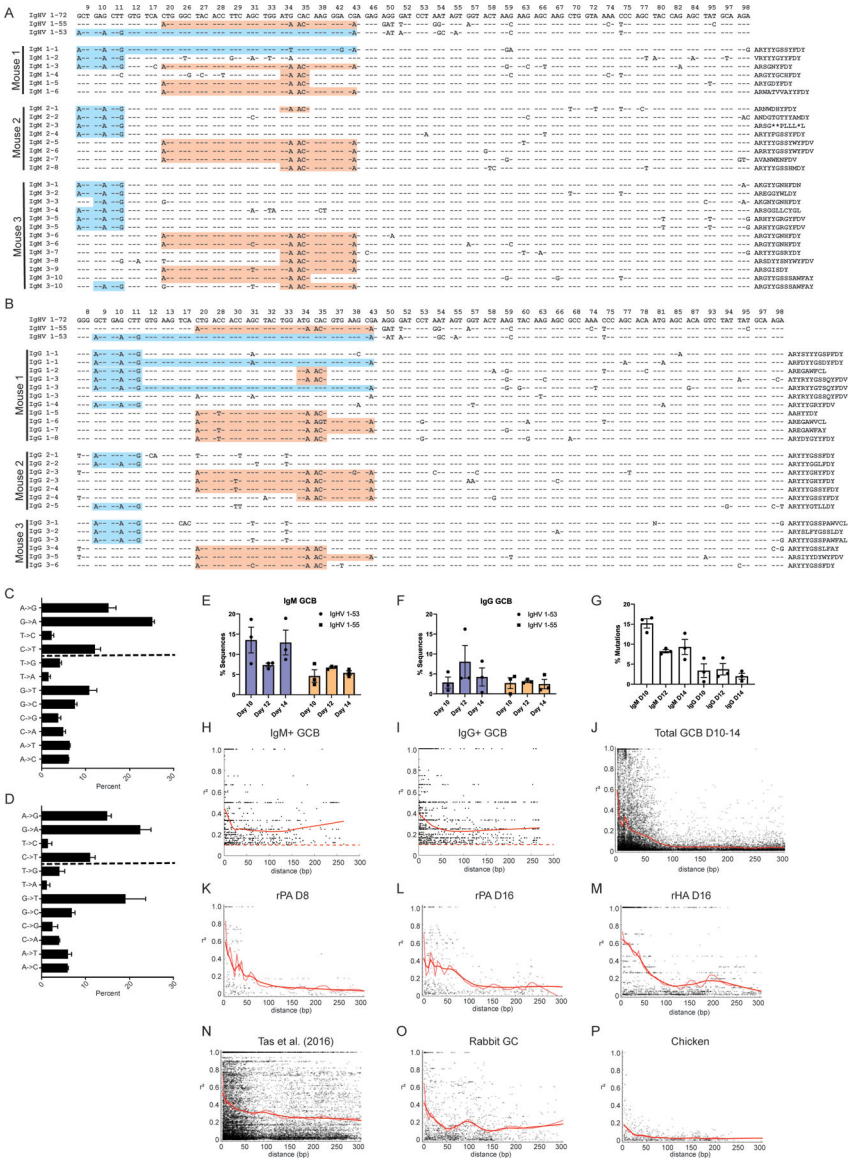

Fig. 2. IgM$^+$ and IgG$^+$ germinal center B cells exhibit gene conversion tracts during the germinal center reaction.

(**A**) Nucleotide alignments of IgM$^+$ germinal center B cells. Data is depicted as in Figure 1C. Sequences are named such that the first number corresponds to the source animal and the second corresponds to unique clones. (**B**) Nucleotide alignments of IgG$^+$ germinal center B cells. Data is depicted as in (**A**). (**C-D**) Percent transition and transversion mutations in IgM (**C**) and IgG (**D**) IgHV 1–72 sequences from day 12 germinal center B cells. Transition mutations are shown above the dashed line, whereas transversions are shown below. (**E**) Shown are the percent of IgHV 1–72 IgM germinal center B sequences in Figure 2A that possess either the IgHV 1–53 tract or IgHV 1–55 tract. (**F**) Shown are the percentage of IgHV 1–72 IgG germinal center B cell sequences in Figure 2B that possess gene conversion tracts as in (**E**). (**G**) Shown are the percent of IgHV 1–72 mutations attributable to gene conversion. (**H-P**) Plot of linkage disequilibrium between all pairs of mutations per sequence per IgHV gene. Data is shown as a function of genetic distance between mutations and

calculated squared correlation coefficient (r^2) of haplotype pairs. The red line represents a LOESS linear regression of the data points. Shown are linkage disequilibrium plots of IgM+ (**H**) and IgG+ (**I**) germinal center B cells with IgHV 1–72 rearrangement; multiple IgHV rearrangements from germinal center B cells at days 10, 12, 14 post-immunization with NPCGG (**J**); day 8 rPA specific germinal center B cells as reported in Kuraoka et al. (2016) (**K**); day 16 rPA specific germinal center B cells (Kuraoka et al., 2016) (**L**); day 16 rHA specific germinal center B cells (Kuraoka et al., 2016) (**M**); Sanger sequenced antibody sequences from multiple IgHV rearrangements as reported by Tas et al. (2016) (n=2150) (**N**);rabbit somatically-mutated IgHV genes (Schiaffella et al., 1999; Sehgal et al., 2002; Winstead et al., 1999) (**O**); and somatically-mutated chicken IgLV segments (Arakawa et al., 2002b; Mansikka et al., 1990) (**P**). For panel H and I, r^2 values are not shown if less than 0.1.

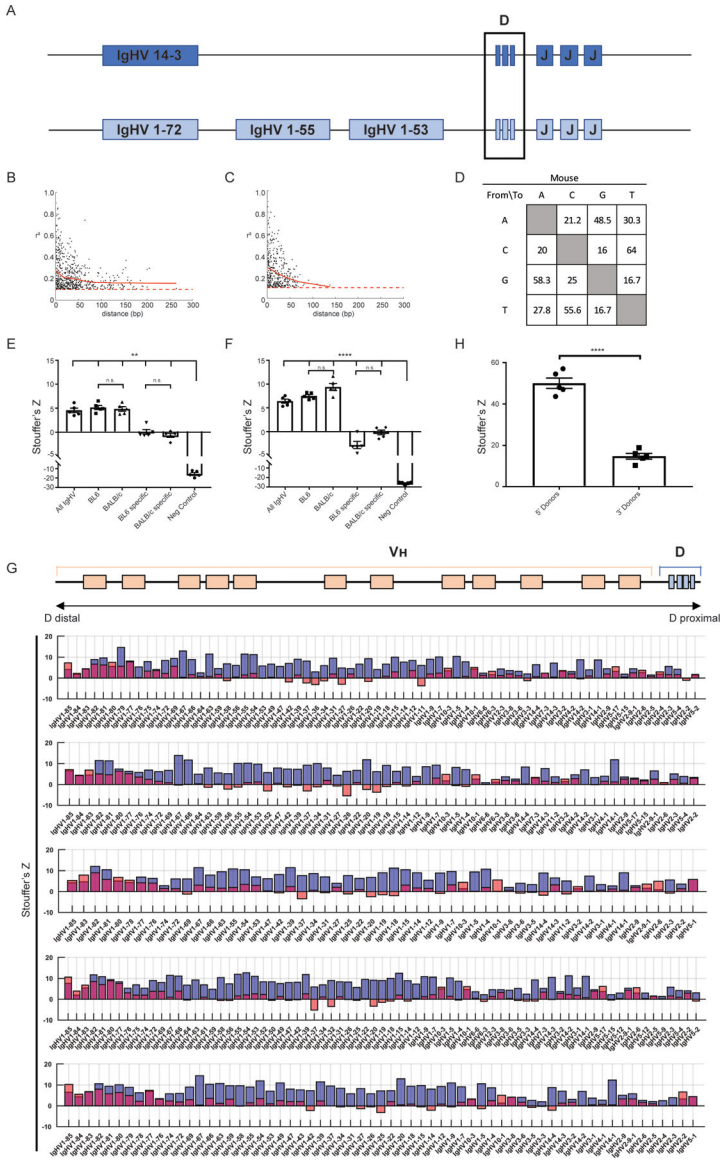

Fig. 3. Germinal center B cells from CB6F1/J mice show templated mutagenesis primarily occurs in cis.

(**A**) Graphic depicting the haplotypes of CB6F1/J mice. Relevant VH exons are shown (**B**) Shown are Stouffer's Z values for PolyMotifFinder and RandomCheck comparisons of somatically-mutated IgHV 1–72 sequences obtained from day 12 germinal centers to different reference sequence sets. All IgHV is the set of all IgHV sequences regardless of strain. BL6 and BALBc refer to IgHV sequences that are specific to each strain, respectively. BL6-specific and BALBc-specific refer to 8-mer motifs that are only present in either BL6 or BALBc mice, respectively. Negative control refers to all 8-mer motifs that are not present in the IgHV repertoire of either strain. (**C**) Data shown as in (**B**) but for somatically-mutated IgHV 14–3 sequences. (**D**) Shown are base pair substitution matrices used for RandomCheck analysis. Tables were obtained from Maul et al. (2016). The table were then transformed from percent of total observed mutations, as reported, to the percent of observed mutations with a given germline nucleotide, such that each row tallies to 100 percent and

indicates the probability of a given base to mutate into another. (**E-F**) Shown are representative r^2 plots of somatically-mutated IgHV 1–72 sequences (**E**) and IgHV 14–3 (**F**). Data is depicted as in Fig. 2. (**G**) A visual schematic of the C57BL/6 IgH locus is shown below which are the Stouffer's Z results of C57BL/6 IgHV rearrangements, grouped by individual CB6F1/J mouse. Each somatically-mutated IgHV is compared against preserved IgHV genes located 5' from the rearranged VDJ or lost IgHV genes 3' to the VH segment that underwent rearrangement. For each IgHV gene, Stouffer's Z against the 5' donors (blue) is overlaid with that of the 3' donors (red). IgHV genes are depicted in the order in which they occur along the IgH locus, with the most D_H-proximal on the right, and the most D_H-distal on the left. (**H**) Stouffer's Z trend is reported for 5' and 3' donors from each mouse. ** $p<0.01$, **** $p<0.0001$, n.s. not significant.

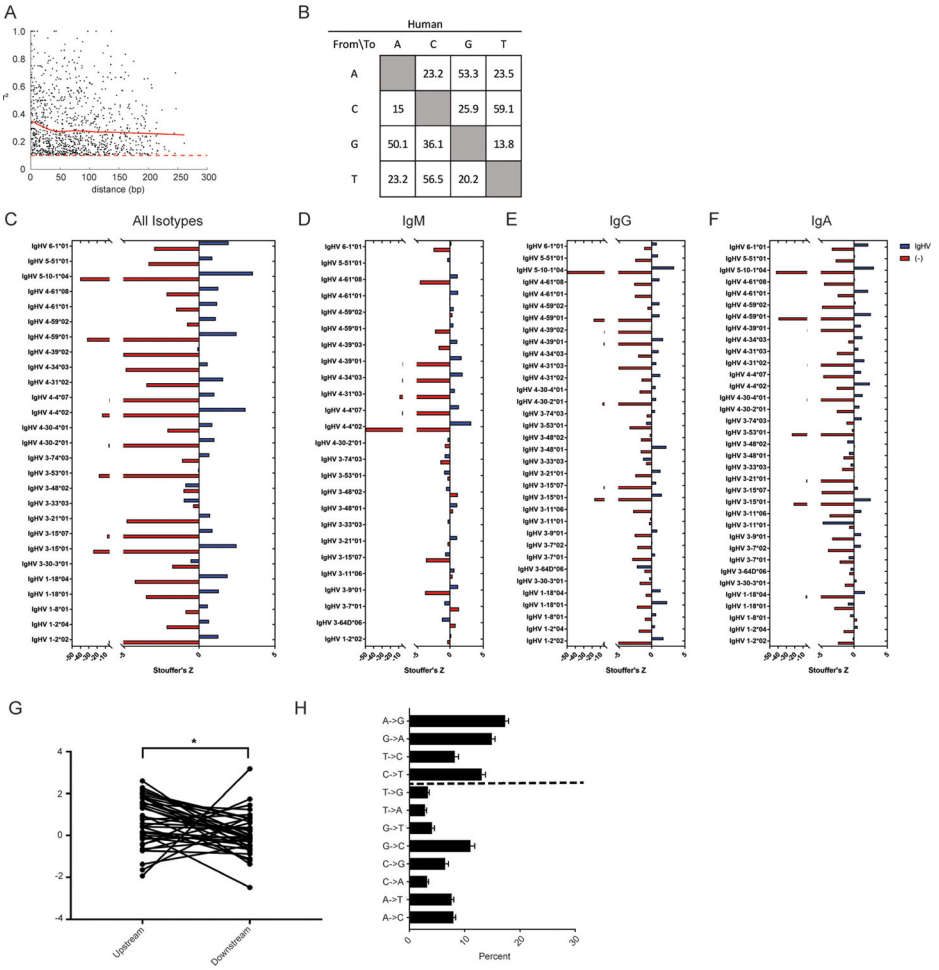

Fig. 4. Human plasmablast sequences demonstrate templated mutagenesis and preferential use of 5' donors.

(**A**) Shown is a r^2 plot of somatically-mutated IgHV 1–2 sequences. (**B**) Base pair substitution matrices used for RandomCheck analysis. Tables were obtained from Longo et al. (2009). Data is presented as in Figure 3G. (**C-F**) Shown are Stouffer's Z scores following PolyMotifFinder/ RandomCheck analysis of somatically-mutated human IgHV genes against either the human IgHV repertoire (blue) or the 8-mer motifs not present in the IgHV repertoire (red) for IgHV genes shared between all isotypes (**C**) or those present of the IgM (**D**), IgG (**E**), or IgA (**F**) isotype. (**G**) Paired dot plot of Stouffer's Z for 5' (upstream) or 3' (downstream) donors for somatically-mutated IgHV sequences. (**H**) Percent transitions and transversions for somatically-mutated IgHV sequences. Data is shown as in Figure 2. *p<0.05

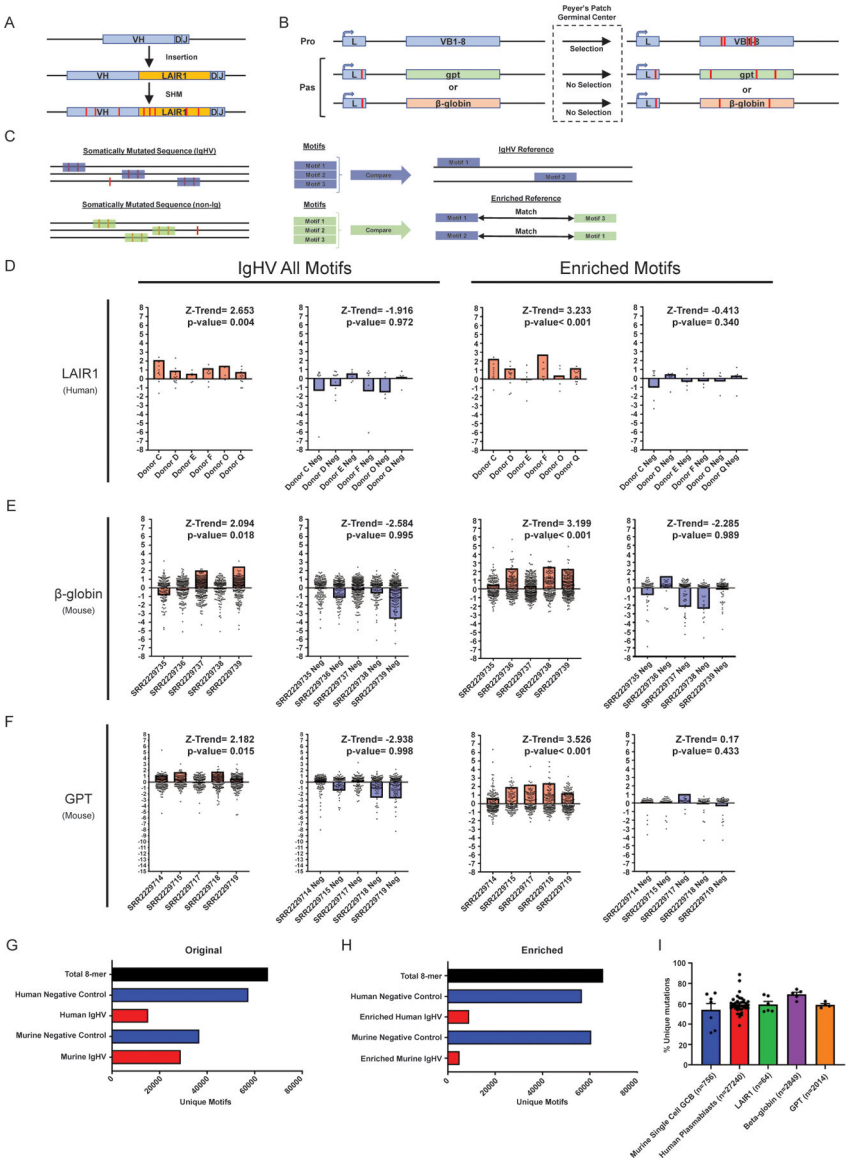

Fig. 5. Non-immunoglobulin sequences exhibit templated mutagenesis and preferentially utilize a limited number of IgHV-specific motifs as donors.

(**A**) Schematic depicting the LAIR1 antibodies described in Tan et al. Somatic mutations are depicted as red bars along the length of the sequence. Only the somatically-mutated LAIR1 segments were used in subsequent analyses (**B**) Schematic depicting the passenger allele transgene system as described in Yeap et al. Leader sequences (L) were intact at the productive allele (pro) but were mutated at the passenger allele (pas) to terminate translation. Only unselected, somatically-mutated gpt or β-globin sequences were used in subsequent analyses. (**C**) Schematic depicting the strategy for enriching motifs used in panels D-F. Somatically-mutated IgHV genes were matched through PolyMotifFinder to IgHV references for each species. In mice, somatically-mutated IgHV 1–72 sequences from CB6F1/J mice were used. In humans, somatic mutations from all IgHVs were used. Motifs produced from two or more mutations within 8bp that matched the IgHV reference were used as a reference for somatically-mutated non-immunoglobulin sequences to be matched

to via PolyMotifFinder/RandomCheck. (**D-F**) Non-immunoglobulin sequence sets were compared via PolyMotifFinder/RandomCheck to either the IgHV repertoire (IgHV all motifs) or to an enriched set of motifs from the IgHV repertoire that were found to be donors somatically-mutated IgHV genes (red). In both analyses, each sequence set is compared to the corresponding number of motifs not in the IgHV set or the enriched set, respectively (blue). Stouffer's Z score is shown as a bar for each analyzed data set and dots represent individual Z scores. For each analysis, Stouffer's Z trend is shown along with the corresponding p-value. Sequence sets shown are LAIR1 (**D**), β-Globin (**E**), and GPT (**F**). (**G**) Shown are the number of motifs used for each analysis in comparison to the total number of unique 8-mers. (**H**) Shown are the number of motifs used in enriched analyses in comparison to the total number of unique 8-mers. (**I**) Shown are the percent of unique mutations that fulfill two conditions: (1) proximity to another mutation within 8bp, and (2) there is a corresponding template present in the IgHV repertoire. GCB data was obtained from the Victora laboratory (Tas et al. *Science* 2016), LAIR1 data was obtained from the Lanzavecchia laboratory (Tan et al. *Nature* 2016; Pieper et al. *Nature* 2017), Beta-globin and GPT data was obtained from the Alt laboratory (Yeap et al. *Cell* 2015). Error bars denote mean ± SEM.

Table 1:

Reagent and Resource Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Alexa Fluor 700 Rat anti-Mouse CD19	BD Biosciences	Catalog No: 557958
Alexa Fluor 647 anti-MU/HU GL7 Antigen	Biolegend	Catalog No: 144605
PerCP-Cy 5.5 Rat anti-Mouse CD45R/B220	BD Biosciences	Catalog No: 561101
BV421 Rat anti-Mouse CD138	BD Biosciences	Catalog No: 562610
APC anti-Mouse CD43	BioLegend	Catalog No: 143208
PE Rat anti-Mouse CD25	BD Biosciences	Catalog No: 558642
PE-Cy7 Rat anti-Mouse IgM	BD Biosciences	Catalog No: 552867
Chemicals, Peptides, and Recombinant Proteins		
$NP_{22}CGG$	Biosearch Technologies	Catalog No: N-5055C-1
Imject Alum	Themo Scientific	Catalog No: 77161
Critical Commercial Assays		
RNeasy Mini Kit	Qiagen	Catalog No: 74104
QIAquick Gel Extraction Kit	Qiagen	Catalog No: 28706
SuperScript III First-Strand	Invitrogen	Catalog No: 18080–051
GS Junior Titanium PicoTiterPlate Kit	Roche	Catalog No: 05996619001
GS Junior Titanium emPCR Kit (Lib-A)	Roche	Catalog No: 05996520001
GS Junior Titanium Sequencing Kit	Roche	Catalog No: 05996554001
Deposited Data		
454 Amplicon Reads	GenBank/SRA	Will upload
Miseq Amplicons Reads	GenBank/SRA	Will upload
Experimental Models: Organisms/Strains		
Mouse: C57BL/6J	The Jackson Laboratory	Stock No: 000664
Mouse: CB6F1/J	The Jackson Laboratory	Stock No: 100007
Sequence-Based Reagents		
Primers for Murine Ig Sequences	**(7)**	N/A
Software and Algorithms		
Matlab (v.R2017b)	MathWorks	https://www.mathworks.com/products/matlab/
Correlation-Based Tests	**(8)**	http://www.niehs.nih.gov/research/resources/software/biostatistics/rxc/index.cfm
Clustal X (v.2.1)	Conway Institute UCD Dublin	http://www.clustal.org/clustal2/#Download
Prism (v.6.0e)	GraphPad	http://www.graphpad.com/scientific-software/prism/
LD-analysis (v.1.0)	This paper	N/A
PolyMotifFinder (v.1.0)	This paper	N/A
RandomCheck (v.1.0)	This paper	N/A

Table depicts the reagents, programs, and strains of mice used for the studies presented.